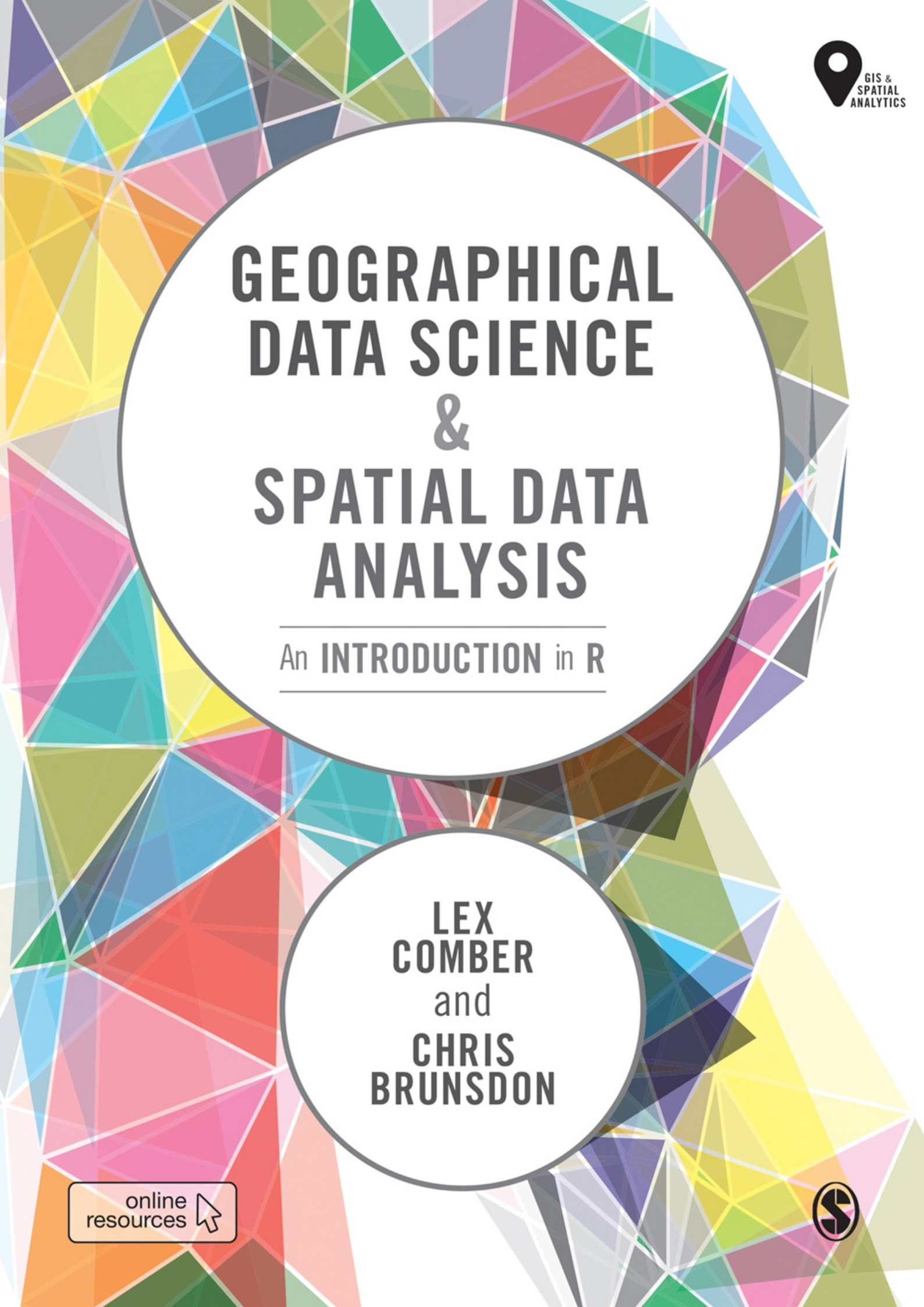
GIS & SPATIAL ANALYTICS
GEOGRAPHICAL DATA SCIENCE & SPATIAL DATA ANALYSIS
An INTRODUCTION in R
LEX COMBER and CHRIS BRUNSDON
online resources

GEOGRAPHICAL DATA SCIENCE & SPATIAL DATA ANALYSIS

An Introduction in R

In the digital age, social and environmental scientists have more spatial data at their fingertips than ever before. But how do we capture this data, analyse and display it, and most importantly, how can it be used to study the world?

Spatial Analytics and GIS is a series of books that deal with potentially tricky technical content in a way that is accessible, usable and useful. Titles include *Urban Analytics* by Alex Singleton, Seth Spielman and David Folch, and *An Introduction to R for Spatial Analysis and Mapping (Second Edition)* by Chris Brunsdon and Lex Comber.

Series Editor: Richard Harris

About the Series Editor

Richard Harris is Professor of Quantitative Social Geography at the School of Geographical Sciences, University of Bristol. He is the lead author on three textbooks about quantitative methods in geography and related disciplines, including *Quantitative Geography: The Basics* (Sage, 2016).

Richard's interests are in the geographies of education and the education of geographers. He is currently Director of the University of Bristol Q-Step Centre, part of a multimillion pound UK initiative to raise quantitative skills training among social science students, and is working with the Royal Geographical Society (with IBG) to support data skills in schools.

Books in this Series:

Geographical Data Science and Spatial Data Analysis, Lex Comber & Chris Brunsdon

An Introduction to R for Spatial Analysis and Mapping, 2nd Edition, Chris Brunsdon & Lex Comber

Agent-Based Modelling and Geographical Information Systems, Andrew Crooks, Nicolas Malleson, Ed Manley & Alison Heppenstall

Urban Analytics, Alex Singleton, Seth Spielman & David Folch

Geocomputation, Chris Brunsdon & Alex Singleton

Published in Association with this Series:

Quantitative Geography, Richard Harris

GEOGRAPHICAL DATA SCIENCE & SPATIAL DATA ANALYSIS

An Introduction in R

Lex Comber
Chris Brunsdon

Los Angeles
London
New Delhi
Singapore
Washington DC
Melbourne

SAGE
Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP
SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320
SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044
SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

First published 2021

Library of Congress Control Number: 2020938055
British Library Cataloguing in Publication data
A catalogue record for this book is available from the British Library
ISBN 978-1-5264-4935-1
ISBN 978-1-5264-4936-8 (pbk)
Editor: Jai Seaman
Assistant editor: Charlotte Bush
Assistant editor, digital: Sunita Patel
Production editor: Katherine Haw
Copyeditor: Richard Leigh
Proofreader: Neville Hankins
Indexer: Martin Hargreaves
Marketing manager: Susheel Gokarakonda

Cover design: Francis Kenney
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in the UK

Lex: To my children, Carmen, Fergus and Madeleine: you are all adults now. May you continue to express your ever growing independence in body, as well as thought.
Chris: To all of my family – living and no longer living.

CONTENTS

ABOUT THE AUTHORS

Alexis Comber

(Lex) is Professor of Spatial Data Analytics at Leeds Institute for Data Analytics (LIDA), University of Leeds. He worked previously at the University of Leicester where he held a chair in Geographical Information Sciences. His first degree was in Plant and Crop Science at the University of Nottingham and he completed a PhD in Computer Science at the Macaulay Institute, Aberdeen (now the James Hutton Institute), and the University of Aberdeen. This developed expert systems for land cover monitoring from satellite imagery and brought him into the world of spatial data, spatial analysis and mapping. Lex's research interests span many different application areas including environment, land cover/land use, demographics, public health, agriculture, bio-energy and accessibility, all of which require multi-disciplinary approaches. His research draws from methods in geocomputation, mathematics, statistics and computer science, and he has extended techniques in operations research/location allocation (what to put where), graph theory (cluster detection in networks), heuristic searches (how to move intelligently through highly dimensional big data), remote sensing (novel approaches for classification), handling divergent data semantics (uncertainty handling, ontologies, text mining) and spatial statistics (quantifying spatial and temporal process heterogeneity). He has co-authored (with Chris Brunsdon) the first 'how to' book for spatial analysis and mapping in R, the open source statistical software, now in its second edition (https://uk.sagepub.com/en-gb/eur/an-introduction-to-r-for-spatial-analysis-and-mapping/book258267). Outside of academic work and in no particular order, Lex enjoys his vegetable garden, walking the dog and playing pinball (he is the proud owner of a 1981 Bally Eight Ball Deluxe).

Chris Brunsdon

is Professor of Geocomputation and Director of the National Centre for Geocomputation at the National University of Ireland, Maynooth, having worked previously in the Universities of Newcastle, Glamorgan, Leicester and Liverpool, variously in departments focusing on both geography and computing. He has interests that span both of these disciplines, including spatial statistics, geographical information science, and exploratory spatial data analysis, and in particular the application of these ideas to crime pattern analysis, the modelling of house prices, medical and health geography and the analysis of land use data. He was one of the originators of the technique of geographically weighted regression (GWR). He has extensive experience of programming in R, going back to the late 1990s, and has developed a number of R packages

which are currently available on CRAN, the Comprehensive R Archive Network. He is an advocate of free and open source software, and in particular the use of reproducible research methods, and has contributed to a large number of workshops on the use of R and of GWR in a number of countries, including the UK, Ireland, Japan, Canada, the USA, the Czech Republic and Australia. When not involved in academic work he enjoys running, collecting clocks and watches, and cooking – the last of these probably cancelling out the benefits of the first.

PREFACE

Data and data science are emerging (or have emerged) as a dominant activity in many disciplines that now recognise the need for empirical evidence to support decision-making (although at the time of writing in the UK at the end of April 2020, this is not obvious). All data are spatial data – they are collected some*where* – and location cannot be treated as just another variable in most statistical models. And because of the ever growing volumes of (spatial) data, from increasingly diverse sources, describing all kinds of phenomena and processes, being able to develop approaches and models grounded in *spatial* data analytics is increasingly important. This books pulls together and links lessons from general data science to those from quantitative geography, which have been developed and applied over many years.

In fact, the practices and methods of data science, if framed as being a more recent term for statistical analysis, and spatial data science, viewed as being grounded in geographical information systems and science, are far from new. A review of the developments in these fields would suggest that the ideas of data analytics have arisen as a gradual evolution. One interesting facet of this domain is the importance of spatial considerations, particularly in marketing, where handling locational data has been a long-standing core activity. The result is that geographical information scientists and quantitative geographers are now leading many data science activities – consider the background of key players at the Alan Turing Institute, for example. Leadership is needed from this group in order to ensure that lessons learned and experiences gained are shared and disseminated. A typical example of this is the *modifiable areal unit problem*, which in brief posits that statistical distributions, relationships and trends exhibit very different properties when the same data are aggregated or combined over different areal units, and at different spatial scales. It describes the process of distortion in calculations and differences in outcomes caused by changes zoning and scales. This applies to **all** analyses of spatial data – and has universal consequences, but is typically unacknowledged by research in non-geographical domains using spatial data.

This possibility of distortion also underpins another motivation for this book at this time: one of *reproducibility*. The background to R, the open source statistical package, is well documented and a number of resources have been published that cover recent developments in the context of spatial data and spatial analysis in R (including our other offering in this arena: Brunsdon and Comber, 2018). This has promoted the notion of the need for *open* coding environments within which analysis takes place, thereby allowing

(spatial) data science cultures to flourish. And in turn this has resulted in a de facto way of working that embraces open thinking, open working, sharing, open collaboration, and, ultimately, reproducibility and transparency in research and analysis. This has been massively supported by the RStudio integrated development environment for working in R, particularly the inclusion of RMarkdown which allows users to embed code, analysis and data within a single document, as well as the author's interpretation of the results. This is truly the 'holy grail' of scientific publishing!

A further driver for writing this book is to promote notions of *critical data science*. Through the various examples and illustration in the book, we have sought to show how different answers/results (and therefore understandings and predictions) can be generated by very small and subtle changes to models, either through the selection of the machine learning algorithm, the scale of the data used or the choice of the input variables. Thus we reject *plug and play* data science, we reject the idea of theory-free analyses, we reject data mining, all of which abrogate inferential responsibility through philosophies grounded in *letting the data speak*. Many of the new forms of data that are increasingly available to the analyst are not objective (this is especially the case for what has sometimes been called 'big data'). They are often collected without any experimental design, have many inherent biases and omissions, and without careful consideration can result in erroneous inference and poor decision-making. Thus being *critical* means considering the technological, social and economic origins of data, including their creation and deployment, as well as the properties of the data relative to the intended analysis, or the consequences of any analysis. Criticality involves thinking about the common good, social contexts, using data responsibly, and even considering how your work could be used in the wrong way or the results misinterpreted. There is no excuse for number crunchers who fail to be critical in their data analysis.

In summary, we believe that the practice of data analytics (actually *spatial* data analytics) should be done in an open and reproducible way, it should include a critical approach to the broader issues surrounding the data, their analysis and consideration of how they will be used, and it should be done wearing geography goggles to highlight the impacts of scale and zonation on the results of analyses of spatial data. This may involve some detective work to understand the impacts of data and analysis choices on the findings – this too is a part of data science. We believe that the contents of this book, and the various coded examples, provide the reader with an implicit grounding in these issues.

REFERENCE